Explainability for Large Language Models: A Survey

Year: 2023 | *Citations: 668* | *Authors: Haiyan Zhao, Hanjie Chen, F. Yang*

Abstract

Large language models (LLMs) have demonstrated impressive capabilities in natural language processing. However, their internal mechanisms are still unclear and this lack of transparency poses unwanted risks for downstream applications. Therefore, understanding and explaining these models is crucial for elucidating their behaviors, limitations, and social impacts. In this article, we introduce a taxonomy of explainability techniques and provide a structured overview of methods for explaining Transformer-based language models. We categorize techniques based on the training paradigms of LLMs: traditional fine-tuning-based paradigm and prompting-based paradigm. For each paradigm, we summarize the goals and dominant approaches for generating local explanations of individual predictions and global explanations of overall model knowledge. We also discuss metrics for evaluating generated explanations and discuss how explanations can be leveraged to debug models and improve performance. Lastly, we examine key challenges and emerging opportunities for explanation techniques in the era of LLMs in comparison to conventional deep learning models.